Audio, visual, and audio-visual egocentric distance perception by moving participants in virtual environments

Marc Rébillat, Xavier Boutillon, Étienne Corteel, Brian F. G. Katz

Audio, visual, and audio-visual egocentric distance perception by moving participants in virtual environments

Marc Rébillat*, Xavier Boutillon†, Étienne Corteel‡, Brian F.G. Katz§

Accepted by ACM Transactions on Applied Perception on 7/25/2012

Abstract

A study on *audio*, *visual*, and *audio-visual* egocentric distance perception by moving participants in *virtual* environments is presented. Audio-visual rendering is provided using tracked passive visual stereoscopy and acoustic wave field synthesis (WFS). Distances are estimated using indirect blind-walking (triangulation) under each rendering condition. Experimental results show that distances perceived in the virtual environment are accurately estimated or overestimated for rendered distances closer than the position of the audio-visual rendering system and underestimated for distances farther. Interestingly, participants perceived each virtual object at a modality-independent distance when using the audio modality, the visual modality, or the combination of both. Results show WFS capable of synthesizing perceptually meaningful sound-fields in terms of distance. Dynamic audio-visual cues were used by participants when estimating the distances in the virtual world. Moving may have provided participants with a better visual distance perception of close distances than if they were static. No correlation between the feeling of presence and the visual distance underestimation has been found. To explain the observed perceptual distance compression, it is proposed that, due to conflicting distance cues, the audio-visual rendering system physically *anchors* the virtual world to the real world. Virtual objects are thus attracted by the physical audio-visual rendering system.

Keywords

Virtual environments, large-screen immersive displays, wave field synthesis, spatialized audio, distance estimation, spatial perception

1 Introduction

Virtual reality (VR) systems aim at providing participants with a virtual world where they would behave and learn as if they were in the real world [10]. Audio-visual (AV) VR-systems that combine large

*LIMSI (UPR CNRS 3251), Université Paris-Sud, 91403 Orsay Cedex, France & LMS (UMR CNRS 7649), École Polytechnique, 91128 Palaiseau Cedex, France

†LMS (UMR CNRS 7649), École Polytechnique, 91128 Palaiseau Cedex, France

‡*sonic emotion labs*, 75015 Paris, France

§LIMSI (UPR CNRS 3251), Université Paris-Sud, 91403 Orsay Cedex, France

immersive screens and many loudspeakers have been developed to provide participants with a virtual space coherently merging holophonic spatial audio and 3D visual renderings [16, 39, 32]. The term *"holophonic spatial audio"* stands here for technologies such as wave-field synthesis (WFS) [6], Ambisonics [18], or others [25], that attempt to physically recreate the same sound field that a real sound source would have radiated, and thus provide participants with a natural spatialized sound rendering. Such AV VR-systems are very appealing because they are minimally intrusive (no headphones needed, only lightweight glasses) and allow participants to move freely in the rendering area while always having a correct AV perspective. With the emergence of these multimodal systems arises the question of the correct perception of the virtual space by moving participants and, more specifically, of rendered distances within it [27, 22].

1.1 Measurement protocols for the estimation of perceived egocentric distance

Because distance perception is a cognitive task, measurement protocols are needed to estimate perceived absolute egocentric distances. Existing measurement protocols can be divided into three main classes [24, 19]: verbal estimations, perceptually directed actions, and imagined actions. In *verbal estimation* protocols, participants assess the perceived distance in terms of familiar units, such as meters. In *perceptually directed action* protocols, an object is presented to the participant who then has to perform an action, such as blind-walking, without perceiving the object any more. In *imagined action* protocols, the action is imagined instead of being performed and response times are used to infer the results of the action. The advantage of *perceptually directed actions* is that they lead to distance estimations that are more accurate and less variable than distance estimations provided by verbal reports [17, 26, 34, 2]. Moreover, using *perceptually directed actions*, estimated distances can be directly inferred from actions whereas a potential systematic bias exists in distances estimated using *imagined action* protocols due to the conversion of a directly measured value of time into an indirect measure of estimated distance [19]. *Perceptually directed actions* have thus been preferred in the present study.

Among perceptually directed actions, *direct* blind-walking and *indirect* blind-walking (triangulation) are two possible alternatives which both lead to accurate distance estimations [17, 26]. Due to physical spatial constraints imposed by the presence of large screens and many loudspeakers, only indirect blind walking (triangulation) is possible in the kind of AV VR-systems under study here [24]. An advantage of the triangulation measurement protocol is that it is applicable to the measurement of audio, visual, and audio-visual perceived absolute egocentric distances without any need to adapt the procedure to each different modality. One disadvantage is that small errors in pointing can lead to large differences in indicated distance for very distant targets. Furthermore, the error is not symmetric since one degree of rotation in one direction can equate to a smaller change in linear distance than an equivalent rotation in the opposite direction.

1.2 Perceived distance in the visual and auditory modalities in real or virtual environments

In classical *visual* VR-systems, such as head-mounted displays (HMD), perceived visual distances have been observed to be systematically underestimated [27, 22]. This is not the case in the real world [41]. VR-systems based on large immersive screens were thought to offer a better distance perception [30]. Studies focusing on visual distance perception in virtual environments rendered by large immersive screens have found that visual distances were underestimated using these systems, exactly as in HMD systems [3, 28, 24, 19, 1].

In the *audio* real world, it is well established that near-auditory distances ($<$ 2m) are overestimated whereas far-auditory distances ($>$2 m) are underestimated (see [42] for a review). Much less is known regarding auditory distance perception in virtual auditory systems based on holophonic spatial audio. In [11, 32], it was shown that holophonic spatial sound renderings can effectively be used to render distances for static sources with moving participants and that perceived distances are compressed with respect to rendered distances. When participants are static, [25, 23] showed that performances in an holophonic audio virtual environment matched well with real world performances in terms of distance perception.

In *audio-visual* virtual environments, perceived visual distances appear to be underestimated, near-auditory distances to be overestimated, and far-auditory distances to be underestimated. Audio and visual perceived distances are thus a priori inconsistent for a given rendered distance. Some efforts have

been done to study how audio and visual distance cues are merged together in virtual environments [14]. Obtained results suggest that static participants perceived AV distances similarly to visual distances. However, participants are rarely static when immersed in virtual worlds. It is thus important to study how AV distances are perceived by participants taking benefit of static and dynamic AV distances cues in a virtual environment.

Furthermore, to provide participants with a virtual world where they would behave as if they were in the real world, VR-systems should be fully *"transparent"* to participants. Transparency is understood here as *"the extent to which the computer displays are capable of delivering an inclusive, extensive, surrounding, and vivid illusion of reality to the senses of a human participant"* [37]. AV VR-systems are however not perfect and suffer from some drawbacks that potentially limit their transparency. It is thus important to assess whether this limitation has an influence on the AV virtual space perceived by participants, and if there exists a spatial link created by AV VR-systems between the real world and the virtual world.

1.3 Objectives

In this paper a study of *audio* (A), *visual* (V), and *audio-visual* (AV) egocentric distance perception in the action space (1.5 m to 6 m) by moving participants in *virtual* environments is presented. AV rendering is provided via the SMART-I^2 platform (Spatial Multi-user Audio-visual Real-Time Interactive Interface) [32, 33] using tracked passive visual stereoscopy and acoustic wave field synthesis (WFS). This AV VR-system allows participants to move freely in the rendering area and maintains stable AV perspective cues everywhere in this area. Distances are estimated by means of perceptually directed action (indirect blind walking, triangulation) under A, V, and AV conditions. This experiment aims at studying how A, V, and AV distances are perceived by participants taking benefit of static and dynamic AV distance cues in a virtual environment. A second objective is to assess whether the lack of total transparency of the AV rendering system induces a spatial tethering between the real and virtual worlds.

2 Method

2.1 Experimental design

To study how A, V, and AV distances are perceived by participants taking benefit of static and dynamic AV distances cues in a virtual environment, five virtual objects (denoted A, B, C, D, E) placed in the participant's *action space*, *i.e.* the space where one "*moves quickly, talks, and if needed can throw something to a compatriot or at an animal*" [15], were rendered (see Fig. 1). To assess whether the position of the AV rendering setup has an influence on the AV virtual space perceived by participants, two initial or starting positions for participants were tested: *Position 1* where participants stood 2.3 m in front of the right panel of the SMART-I^2, and *Position 2* where they stood 3.3 m from it (see Fig. 1). Virtual objects are at the same locations with respect to the rendering system for both starting positions. Virtual objects were located at distances of 1.5 m, 2 m, 2.5 m, 3.5 m, and 5 m from *Position 1*, equating to distances of 2.5 m, 3 m, 3.5 m, 4.5 m, and 6 m from *Position 2*.

A total of 40 volunteers (30 men, 10 women) between 21 and 49 years old participated in the experiment with half of participants starting from *Position 1* and the other half starting from *Position 2*. All participants had self-reported normal vision (possibly corrected) and normal hearing. Each participant had to estimate the distances of the five virtual objects four times under each rendering condition. They performed three sessions of 20 iterations each after a training phase of two iterations under each rendering condition. In the training phase, rendered distances were 3 m and 7 m for *Position 1*, and 4 m and 8 m for *Position 2*. Participants took pauses between sessions and the entire experiment lasted approximately one hour. The session order was balanced between the six possible permutations of the three rendering conditions.

The chosen experimental design was therefore a mixed design with three factors: rendered distance d_r (five levels, within-participants), rendering condition (three levels, within-participants), and starting position (two levels, between-participants). The dependent variables are perceived distance d_p, time t_{XP} spent in the exploration phase (see Sec. 2.4), and exploration path length l_{XP}.

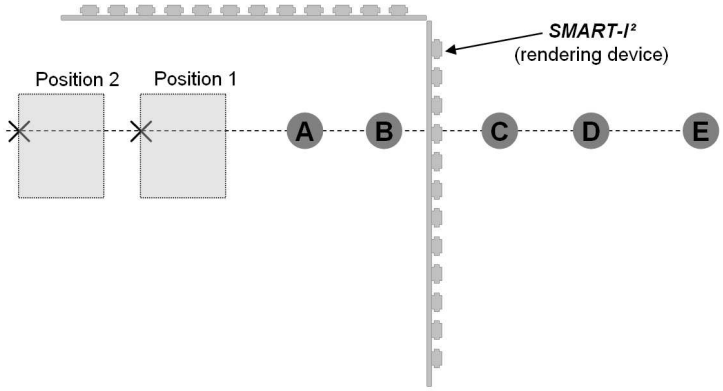


Figure 1: Overview of the experimental setup. *Virtual objects*: grey disks labelled A, B, C, D or E. *Start positions*: black «×». *Exploration areas* are represented by the grey rectangles

2.2 Experimental setup

Experiments were conducted in the AV virtual environment produced by the SMART-I^2 platform [32, 33]. In this system, front-projection screens and loudspeakers are integrated together to form large flat multi-channel loudspeakers also called *Large Multi-Actuator Panels* (LaMAPs). The rendering screens consist of two LaMAPs (2 m × 2.6 m, each supporting 12 loudpeakers) forming a corner (see Fig. 2). The reporting interface used in the present experiment was a *wiimote*.

Visual rendering was produced using tracked passive stereoscopy rendered at 80 frames per second with a resolution of 1280 × 960 pixels on each screen. Interocular distance for stereoscopic rendering was fixed at 6 cm for all participants. At both starting positions (the black «×» in Figs. 1, 3(a), and 3(b)), the horizontal field of view was approximately 150 ° and the vertical field of view approximately 70 ° . Since it has been shown that graphical resolution [35, 19] and field of view [13] have no influence on *visual* distance perception, these experimental parameters should not influence the obtained experimental results.

Spatial audio rendering was realized via acoustic Wave Field Synthesis (WFS) [6]. This technology attempts to physically recreate the acoustic sound field corresponding to a virtual source at any given position in the horizontal plane, without the need for tracking. Real-time audio signal processing was achieved by a *Wave 1* rendering engine provided by *sonic emotion*. The inter-loudspeaker distance of 21 cm corresponds to an aliasing frequency $f_{al} \simeq 1.1$ kHz, up to which the sound field is correctly reconstructed [11]. It has been demonstrated by [36] that sound fields reconstructed by WFS are sufficiently consistent to allow for accurate localization, even when frequencies above f_{al} are present. In [12], it was shown that even if not *exact*, azimuthal cues above the aliasing frequency f_{al} are generally consistent with azimuthal cues below f_{al} when using MAPs.

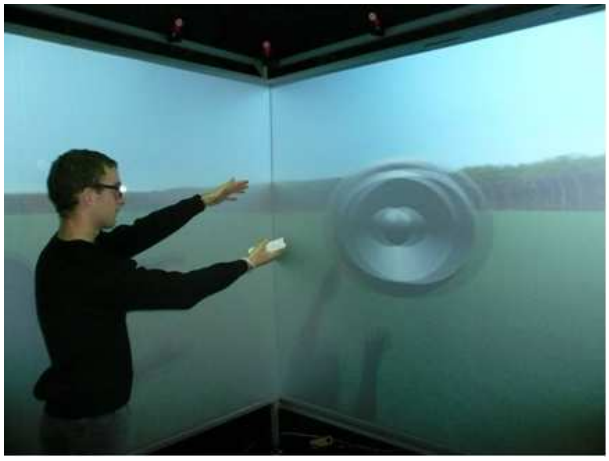
Furthermore, fine temporal and spatial calibration has been performed to ensure that the audio and visual renderings are fully coherent.

2.3 Audio, visual, and audio-visual stimuli

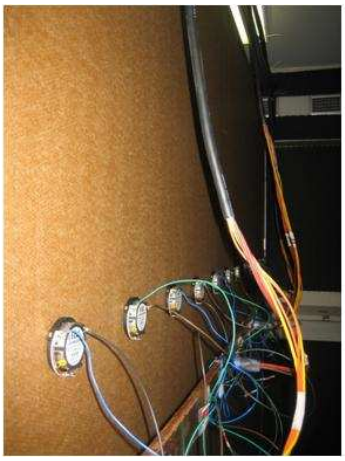
The *visual environment* consisted of an open, grassy field with a forest at 50 m (Fig. 2(a), trees were $\simeq 7$ m tall). The associated *audio environment* consisted of the sound of wind in the trees accompanied by some distant bird songs (overall background level of 36 dBA). The audio environment was created by 12 plane waves equally distributed in the horizontal frontal field of rendering (*i.e.* between -70 ° and 70 °). Environmental sound levels were adjusted to be slightly above the background noise produced by the video-projectors (background noise level of 34 dBA).

The chosen *visual target* object was a footless 3D loudspeaker, approximately spherical, with a diameter of $\simeq 30$ cm (Fig. 2(a)). The stand was removed to avoid window violations when the object was displayed in front of the screen. The floating loudspeaker was positioned at a height of 1.6 m and shadows were displayed.

The associated *audio target* object was a 4 kHz low-pass filtered white noise with a 15 Hz amplitude modulation. Low-pass filtered white noise has been chosen in order to have a wide spectral content and



(a) Front view



(b) Back view

Figure 2: *Left*: a participant and an audio-visual object in the virtual world provided by the SMART-I^2. Visual rendering is projected on the front faces of the two LaMAPs which form a corner. *Right*: electro-dynamical exciters are glued on the back of each LaMAP.

to allow participants to rely on numerous audio localization cues. The white noise was modulated in amplitude by a sine wave to produce attack transients that are also useful in sound localization [7]. No simulated room-effect (*i.e.* ground reflection) was included. The sound level of the omnidirectional audio object corresponds to a monopole emitting 78 dB(SPL) at 1 m, well above the environmental sound level at each of the tested distances.

Audio and *visual* objects were always displayed coherently, *i.e.* at the same spatial position. In addition, their *visual* size and *audio* level decreased naturally with distance. As the experimental design allowed participants to move within the rendering area (see Sec. 2.4), they could rely on a large number of cues naturally available in the corresponding real environment for the estimation of distances, including dynamic cues. In particular, *motion parallax*, which denotes changes in the angular direction of a point source occasioned by the participant's translation is available. This cue has been shown to be useful for distance estimation using the visual [5, 29] or the auditory modality [38, 31]. Another dynamic cue, the estimated time-to-impact for a constant velocity between the moving participant and the static source (also denoted acoustic or visual τ), can also be used [4, 31]. Available AV distance cues are summarized in Tab. 1.

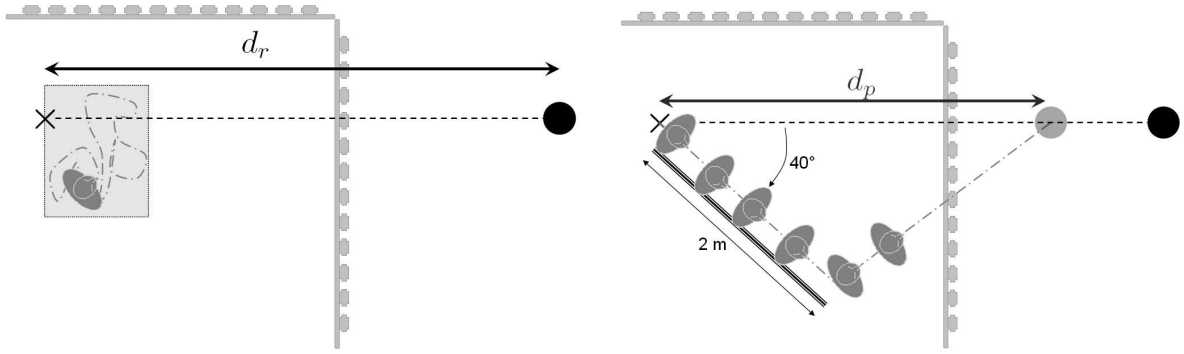
Available cue	Modality	Class
Object size/level	A, V	Relative*
Motion parallax	A, V	Absolute
Time-to-impact	A, V	Absolute
Binocular/binaural cues	A, V	Absolute†
Height in the visual field	V	Relative*

Table 1: Available audio-visual cues.

The AV background environment was kept active in all the rendering conditions. In the *audio* condition, the spatialized sound corresponding to the virtual object was played while no image of the virtual object was shown. The only visual image consisted of the open, grassy field with a forest in the background. In the *visual* condition, the 3D image of the virtual object was displayed with no corresponding sound. The only audio signal consisted of the sound of wind in the trees accompanied by some bird songs. In the *audio-visual* condition, the spatialized sound of the virtual object was rendered with its corresponding 3D image and the AV environment.

2.4 Experimental task

Distance estimation was performed in this task in two phases: a *presentation* phase, see Fig. 3(a), and a *reporting* phase, see Fig. 3(b). Participants began each iteration at one of the two possible *start positions*, indicated by a black «×» in Figs. 1, 3(a), and 3(b).



(a) Presentation phase *(b)* Reporting phase

Figure 3: Presentation and reporting phases. *Start position*: «×». *Virtual object*: black disk placed at a rendered distance d_r from the *start position*. In Fig. 3(a), the *exploration area* is represented by the grey rectangle and the dotted grey line indicates a typical exploration trajectory performed by the participant. In Fig. 3(b), the *guide* is shown as a thick plain black line. The dotted grey line indicates a typical trajectory performed by the participant. *Perceived object*: grey disk placed at the estimated perceived distance d_p from the *start position*.

Before starting the *presentation* phase, participants had to indicate that they were ready to perform this phase by pressing a *wiimote* button. In the *presentation* phase, participants moved around in the *exploration* area which was a rectangle of 1.0×0.8 m^2. Participants were instructed to move in the *exploration* area in order to acquire "*a good mental representation of the virtual object and its environment.*" A typical path followed by a participant during the *presentation* phase is depicted in Fig. 3(a).

Once a "*a good mental representation*" has been acquired, participants pressed a button indicating that they were ready for the *reporting* phase. At this point, the target stimuli was stopped, and the procedure for distance estimation by means of triangulated blind walking began, as depicted in Fig. 3(b). Participants closed their eyes, made a 40 ° right-turn to a handrail guide which was included to help during blind-walking, and walked blindly for an imposed distance of $\simeq 2$ m, following the handrail to the end. Participants stopped at the end of the guide, turned in the direction where they thought the object was, and took a step forward in the direction of the source position. Participants had been instructed that the perceived distance was to be calculated according to this step. They then indicated that they had completed the reporting phase by again pressing a button. Afterwards, they could open their eyes and return to their initial *start position* for the next trial. The experimental protocol was fully automated, with the participants being observed remotely so as not to disturb the sense of presence.

2.5 Post-session questionnaire

In the present experiment, participants were asked to complete a 7-item questionnaire at the end of each of the three experimental sessions (A, V, AV). The goal of this questionnaire was to evaluate the feeling of *presence* that participants experienced during each session. This questionnaire was built by adapting statements taken from [8] and [3], translated into French. Statements were rated on a 7-point Likert scale ranging from -3 to 3 with two anchors. The statements are provided in Tab. 2.

3 Analysis of results

3.1 Extraction of the dependent variables

This section explains how the different dependent variables (d_r, t_{XP}, l_{XP}) were derived from the experimental data. The position of the head of the participant (central point between the eyes) is

Q1†:	I had the feeling of locating a real object.
Q2†:	I had the feeling of looking at a TV instead of really being in an outdoor environment.
Q3†:	The virtual environment became real for me and I forgot the real environment.
Q4†:	I remember the virtual environment more as a place where I have been than as a computer generated image I have seen.
Q5*:	I had the impression that I could touch the virtual objects.
Q6*:	I felt present in the virtual world.
Q7†:	I felt surrounded by the virtual world.

Table 2: Post-session questionnaire.. †: Statements adapted from [8]. *:Statements adapted from [3].

recorded for each iteration during both the *presentation* and *reporting* phases at 100 Hz.

The duration of the *presentation* phase t_{XP} was obtained by measuring the time between the participant's button presses for *ready for a new trial* and for *ready for the reporting phase*, as explained in Sec. 2.4. The exploration path length l_{XP} was calculated by using the head position recorded during t_{XP}.

The exploration path length walked during the exploration phase l_{XP} can be separated into the component walked parallel to the direction of the source l^P_{XP} and the component walked in the orthogonal direction l^O_{XP}. To be comparable, these two paths were normalized by the maximum physical path lengths in each direction, which are here the sides of the exploration area (*i.e.* $s_O = 1$ m and $s_P = 0.8$ m). The dependent variable $P = l^P_{\mathrm{XP}}/s_P$ (respectively $O = l^O_{\mathrm{XP}}/s_O$) denotes the number of times the participant walked the length of the exploration area parallel (respectively orthogonal) to the source direction.

Perceived distances d_p were estimated from the *triangulation* trajectory as follows: a line ($y = ax + b$) was fitted to the trajectory points during the forward step (118 ± 67 points have been used for the fit). The estimated perceived distance is given by Eq. (1):

$$d_p = -\frac{b}{a} \tag{1}$$

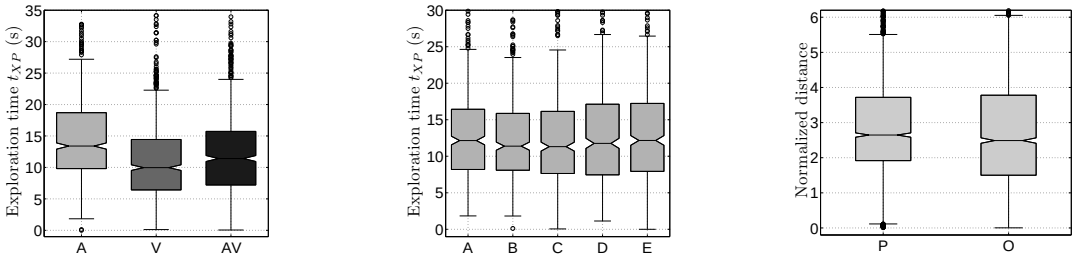
The relative 95%-confidence intervals on the estimated distance are deduced from the 95%-confidence intervals of the linear fit regression coefficients a and b for each iteration (`regress` function in Matlab). Accross all iterations, participants, and distances, the 95%-confidence intervals for the relative distances, *i.e.* for d_p/d_r, estimated using the triangulation trajectory is $\pm 8.5\%$. The triangulation procedure and the associated data treatement thus provide a reliable estimation of the perceived distances.

3.2 Presentation phase

In this section, the influence of the rendered distance d_r and of the condition (A,V,AV) on the time t_{XP} and on the path length l_{XP} respectively spent and walked during the *presentation phase* are analyzed. Data collected for *Position 1* and *Position 2* are pooled together as a one-way ANOVA performed on the exploration time t_{XP} with factor *starting position* showing no significant difference ($F = 1.94$ and $p < 0.17$). Results of the analysis are shown in Figs. 4.

A two-way repeated-measures analysis of variance (ANOVA) performed on the *exploration time* t_{XP} with *condition* (A,V,AV) and *rendered distance* d_r as within-participant factors shows that *condition* is highly significant ($F(2, 64) = 7.87$ and $p < 0.008$), that *rendered distance* is significant ($F(4, 64) = 3.19$ and $p < 0.015$), and that there is no interaction effect between *condition* and *rendered distance* d_r ($F(8, 64) = 0.52$ and $p < 0.84$). Post-hoc tests, computed in terms of medians are shown for *condition* in Fig. 4(a). They reveal that *exploration times* for each *condition* are significantly different. Post-hoc tests for *rendered distances* shown in Fig. 4(b) revealed that *exploration times* for the virtual object C are slightly, but significantly, lower than those obtained for the virtual object A. Participants spent more time in the *exploration phase* when estimating distances using the audio modality than when using the audio-visual modality. Furthermore, participants spent more time in the *exploration phase* when estimating distances using the audio-visual modality than when using only the visual modality.

Normalized exploration path lengths in the direction of the virtual object (P) and in the orthogonal direction (O) are compared in Fig. 4(c). The analysis reveals that P is slightly, but significantly, longer than O. Given a certain exploration area, participants walked 1.06 times longer in the direction parallel to the virtual object than in the direction perpendicular to the virtual object during the exploration phase.

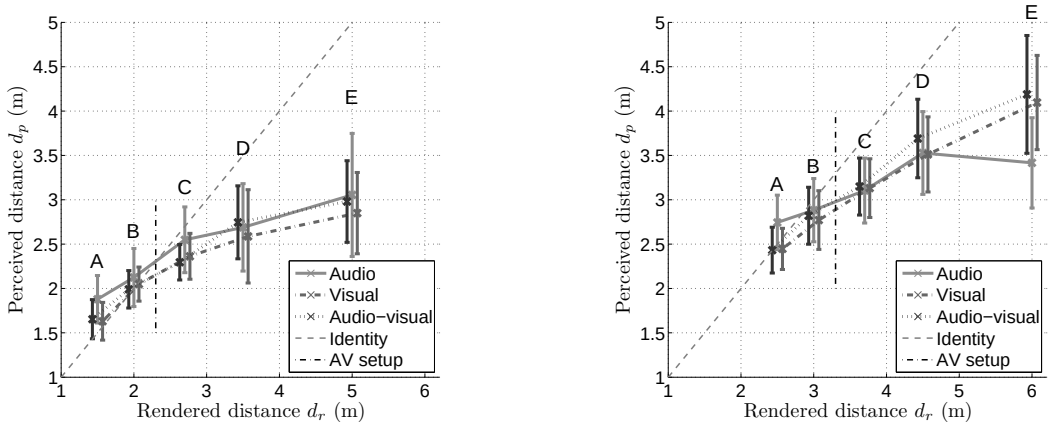


(a) *(b)* *(c)*

Figure 4: Exploration time (t_{XP}) as a function of (a) the condition and (b) virtual objects, and (c) comparison of the normalized exploration path lengths in the direction of the virtual object (P) and in the perpendicular direction (O). *On each box, the central mark is the median, the edges of the box are the* 25^{th} *and* 75^{th} *percentiles, the whiskers extend to the most extreme data points (outliers not considered). Points are drawn as outliers if they are greater than* $q_3 + 1.5(q_3 - q_1)$ *or less than than* $q_1 - 1.5(q_3 - q_1)$*, where* q_1 *and* q_3 *are the* 25^{th} *and* 75^{th} *percentiles, respectively. Notches denote comparison intervals. Two medians are significantly different at the 5% significance level if their intervals do not overlap. Interval endpoints are the extremes of the notches.*

3.3 Reporting phase

In this section, the influence of the rendered distance d_r and of the rendering condition (A,V,AV) on the perceived distances d_p is analysed for each starting position. Means and standard deviations of perceived distances for participants starting from *Position 1* are shown in Fig. 5(a) and for participants starting from *Position 2* in Fig. 5(b).



(a) Position 1 *(b)* Position 2

Figure 5: Mean and standard deviation of perceived distances d_p as a function of rendered distance d_r for each rendering condition and each starting position. Vertical lines represent one standard deviation.

For *Position 1*, the perceived distances d_p were analysed using a repeated-measures two-way ANOVA with *rendered distance* d_r and *rendering condition* (A,V,AV) as factors. *Rendered distance* d_r is significant at the 5% level with $F(4, 64) = 52.93$ and $p < 10^{-6}$. *Rendering condition* is not significant at the 5% level as $F(2, 64) = 2.23$ and $p < 0.12$. No interaction between *rendered distance* d_r and *condition* has been found since $F(8, 64) = 0.95$ and $p < 0.47$. As post-hoc tests, a series of Bonferroni corrected *t-tests* have been performed and all the *rendered distance* pairs have been found to be significantly different.

For *Position 2*, the perceived distances d_p were analysed similarly. *Rendered distance* d_r is significant at the 5% level with $F(4, 64) = 48.79$ and $p < 10^{-6}$. *Rendering condition* is not significant at the 5% level as $F(2, 64) = 1.84$ and $p < 0.17$. A significant interaction is found between *rendered distance* d_r and

condition as $F(8, 64) = 8.95$ and $p < 10^{-6}$. The virtual object A is perceived significantly farther in the A condition than in the V or AV conditions. The virtual object E is perceived significantly closer in the A condition than in the V or AV conditions. As post-hoc tests, a series of Bonferroni corrected *t-tests* have been performed and all the *rendered distance* combinations have been found to be significantly different.

For both starting positions, the different distances d_r are thus correctly ordered and well recognized by participants, independently of the rendering condition. Interestingly, participants perceived each virtual object at a modality-independent distance when using the audio modality, the visual modality, or the combination of both. The audio-visual spatial rendering provided by the SMART-I^2 is, in this sense, fully coherent in distance. By comparing Figs. 5(a) and 5(b), it can furthemore be seen that the starting position has a direct impact on distance perception. This aspect of the results will be discussed in details in Sec. 4.

A very small influence of the presentation order has been observed: participants presented with the audio condition in third position made slightly larger errors than participants presented with the audio condition in the first position. However, this effect remains small. The possibility of any learning effect that could have occurred during the 60 trials of the experiment has been checked by comparing groups of 10 successive trials. No significant differences between the relative errors made by the participants among the different groups of trials have been found. Thus, no learning effect appeared during the experiment.

3.4 Post-session questionnaires

At the end of each session (A, V, and AV), participants rated 7 statements on a 7-point Likert scale with two anchors (see Tab. 2). As differences between the different sessions are to be analyzed for each statement, any bias due to participants has been removed using the following procedure: the rating $A_n^i(k)$ of the n^{th} participant for the k^{th} statements during session i (i = A,V,AV) has been transformed into $\underline{A}_n^i(k) = A_n^i(k) - M_n(k)$, with $M_n(k)$ the mean over the three sessions of the ratings of the n^{th} participant for the k^{th} statements. The *presence*-score has been built as the mean of the unbiased ratings $\underline{A}_n^i(k)$.

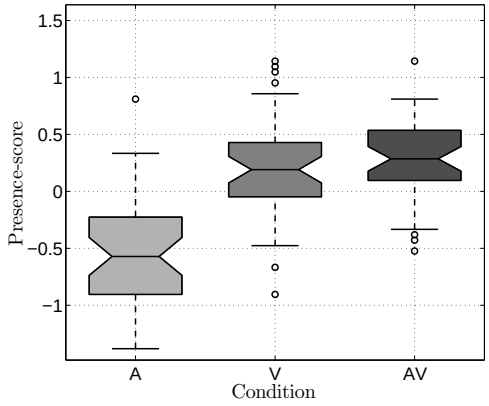


Figure 6: *Presence*-score for each of the rendering condition. For explanation regarding boxplots, see the caption of Fig. 4.

Results shown in Fig. 6 reveal that the scores of the A condition are significantly lower than the scores of the others conditions (V and AV) and that the V and AV conditions are not significantly different. Thus, *presence* is rated significantly lower in the A condition than in the V and AV ones. Moreover, *presence* is rated statistically equivalently for the V and AV rendering conditions.

4 Influence of the presence of the AV VR-system on the perceived virtual space

4.1 Potential conflictual audio-visual spatial cues

Like the vast majority of virtual and augmented reality systems, the SMART-I^2 system is not perfect and potentially provides conflicting audio-visual spatial cues. As specified in Sec. 2.3, no room effect was synthesized in order to recreate acoustical conditions that were as close as possible to open free-field conditions. Nevertheless, even though the experimental room had been acoustically treated, there were still traces of a room effect, with a mid-frequency mean reverberation time T_{60}(500 Hz to 1 kHz) of 0.45 s. The ratio of the energies of the direct and reverberated components of the sound, which is an audio distance cue [9], specifies to the participant a distance corresponding to the physical setup rather than the distance to the virtual object.

The technology used to provide the 3D visual rendering is not perfect either. To estimate the distance of the virtual visual object, participants use two binocular cues. Focus cues (accommodation and blur in the retinal image) specify the distance at which the screen, instead of the virtual object, is seen. Vergence cues correspond to the distance at which the optical axes of the two eyes cross one another, *i.e.* the virtual object. In a 3D visual rendering setup based on large immersive screens, focus cues are then almost always in conflict with vergence cues [21] which can affect depth perception [40, 20]. Finally, shadows projected on the virtual ground floor were visible only for the virtual objects D and E but not for the nearer ones A, B, and C. For close distances, the lack of shadow is thus also in conflict with other spatial cues.

4.2 Anchor hypothesis

The possible presence of conflictual audio-visual cues can potentially have an effect on distance perception. If participants are experiencing audio-visual cues specifying two different distances, it is expected that the virtual object will be perceived somewhere between these two distances. Furthermore, the only distance at which all cues are in agreement corresponds to the physical location of the AV VR-system. Distance perception is thus expected to be correct at that position. To test for the possible existence of such an effect, two starting positions (*Position 1* and *Position 2*) *i.e.* two physical locations of the AV VR-system, have been included in the experimental design (see Sec. 2.1). In the results presented in Figs. 5(a) and 5(b), virtual objects rendered in front of the LaMAP (*i.e.* A and B) appear to be pushed toward the LaMAP whereas virtual objects rendered behind the LaMAP (*i.e.* C, D, and E) seem to be pulled toward it. Furthermore, the distance at which the perceived distance equals the rendered distance corresponds roughly to the distance between the participants and the physical location of the AV VR-system (*i.e.* $D_s^{P1} = 2.3$ m for *Position 1* and $D_s^{P2} = 3.3$ m for *Position 2*). It is hypothesized that, because some audio-visual cues specify the distance to the physical setup instead of the distance of the virtual object, participants tend to bind their perceptual distance estimation to the actual rendering system setup. In that sense, the AV VR-system thus physically anchors the virtual world to the real world. At this point, it must be noted that two distances must be considered during the experiment: the distance that must be *evaluated* by the participant, which is defined explicitly as the distance between the object and the initial position ($P1$ or $P2$) of the participant, and the distance between the object and the participant, which is varying during the exploration phase. The latter is used by the participant to evaluate the former. It is also hypothesized that anchoring, if such an effect exists, pertains to the average of this latter distance, which is *experienced* by the participant during the exploration phase.

Following [42] for the audio modality and [41] for the visual modality, it is assumed that a compressive model in the form $d_p = k \times (d_r)^a$ relates the perceived distance d_p to the rendered distance d_r. The coefficient a denotes the global perceptual compression and is not expected to be influenced by the starting or exploring position of the participants. However, the value of a may differ between the different rendering conditions. If participants are located at a distance D_a from the rendering device, the anchor hypothesis predicts the value of k, and the relation between d_p and d_r should be:

$$d_p = D_a \times \left(\frac{d_r}{D_a}\right)^a \tag{2}$$

The anchor hypothesis thus predicts for each starting position that:

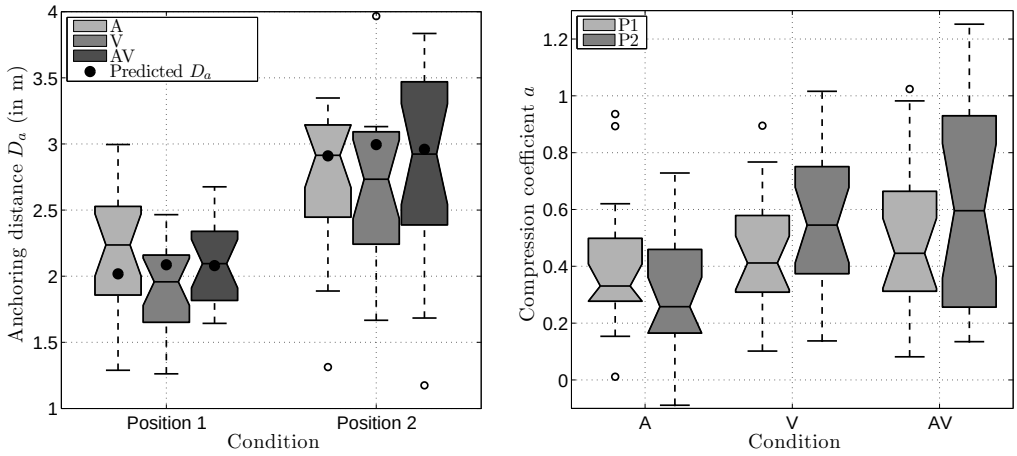
$$\text{Position 1} \quad \rightarrow \quad D_a^{P1} = D_s^{P1} + < \underline{l_{\text{XP}}^P} > \quad \text{with} \quad D_s^{P1} = 2.3 \text{ m}, \quad \text{and} \quad a^{P1} = a \tag{3}$$

$$\text{Position 2} \quad \rightarrow \quad D_a^{P2} = D_s^{P2} + < \underline{l_{\text{XP}}^P} > \quad \text{with} \quad D_s^{P2} = 3.3 \text{ m}, \quad \text{and} \quad a^{P2} = a \tag{4}$$

where $< \underline{l_{\text{XP}}^P} >$ is the average of the algebraic walking displacement of the participant in the direction of the source during the exploration phase. If no correlation is observed between the couple of crossing distances between curves in Fig. 5 (for Position 1 and Position 2) and the physical distances to the screen, the anchor hypothesis does not stand.

4.3 Experimental evidence of the anchor effect

For the two starting positions that have been tested, the values of D_a and a for each participant and for each rendering condition have been estimated by fitting the compressive model of Eq. (2) to the collected data. Among all fits, a mean $R^2 = 75.3$ % is obtained, highlighting the high quality of the model. The estimated and predicted anchoring distances D_a and the compression coefficients a are plotted versus the rendering condition (A, V, AV) and the starting position (*Position 1*, *Position 2*) in Figs. 7.

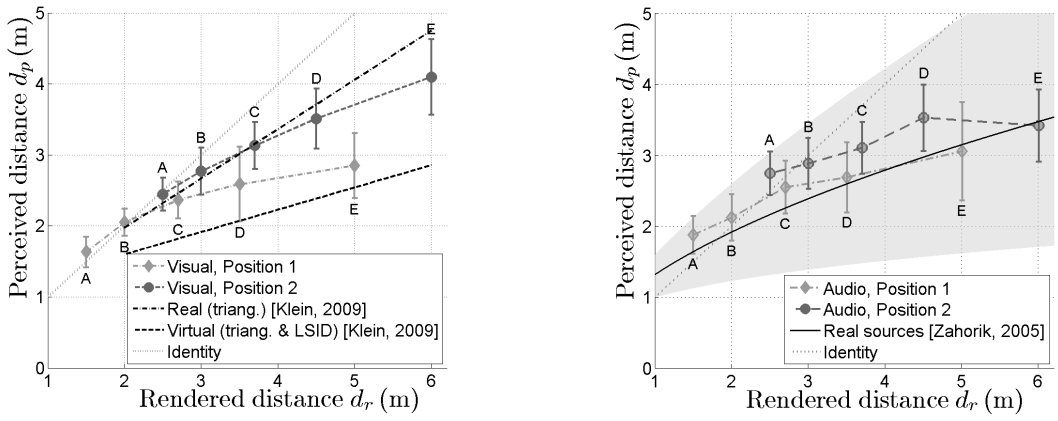


(a) Anchoring distance D_a *(b)* Compression coefficient a

Figure 7: Anchoring distance D_a and compression coefficient a versus starting position (*Position 1*, *Position 2*) and rendering condition (A, V, AV). For explanation regarding boxplots, see the caption of Fig. 4.

From Fig. 7(a), it can be observed that the anchoring distances D_a corresponding to each rendering condition are not significantly different for a given starting position. Moreover, for all rendering conditions the anchoring distances D_a are significantly larger for *Position 2* than for *Position 1*. For the *audio* condition, median values of $D_a^{P1} = 2.24$ m and $D_a^{P2} = 2.91$ m are obtained while Eqs. (3) and(4) predict 2.02 m and 2.90 m respectively, see Fig. 7(a). For the *visual* condition, median values of $D_a^{P1} = 1.96$ m and $D_a^{P2} = 2.73$ m are obtained (*vs.* predicted values of 2.09 m and 2.99 m respectively). For the *audio-visual* condition, median values of $D_a^{P1} = 2.09$ m and $D_a^{P2} = 2.92$ m are obtained (*vs.* predicted values of 2.08 m and 2.96 m respectively). From Fig. 7(b), it can be observed that the compression coefficient a corresponding to each rendering condition is not significantly different for any rendering condition between *Position 1* and *Position 2*. A median value of $a = 0.31$ is obtained for the *audio* condition, with $a = 0.48$ for the *visual* condition, and $a = 0.45$ for the *audio-visual* condition.

The anchor hypothesis predicts, according to Eqs. 3 and 4, that the anchoring distance D_a should be at the positions indicated by the black dots in Fig. 7(a). The quantitative agreement between these predictions and the experimental anchoring distances is excellent, specifically for the AV condition. The anchor hypothesis furthermore predicts that the compression coefficient a should not be different between *Position 1* and *Position 2*. This is effectively the case, as shown by Fig. 7(b). This experimental evidence thus argues in favor of the anchor hypothesis proposed in Sec. 4.2.



(a) Visual modality

(b) Audio modality

Figure 8: Comparison of results obtained for *visual* and *audio* modalities with some results from the literature. For the *visual* modality, [24] studied, using triangulation, egocentric distance perception in an open grassy field, in the real world and in a virtual world rendered by LSID. For the *audio* modality, [42] proposed a psychophysical curve relating perceived distance to real distance from a review covering 84 experiments. The shadowed zone denotes the standard deviation associated with the results from the 84 experiments.

5 General discussion

5.1 Perceived distance in *visual* large screens immersive displays (LSID)

[24] have studied, using triangulation, visual egocentric distance perception in an open grassy field, in both the real world and rendered via LSID in a virtual world. Their results can thus be directly compared to the results obtained here for the *visual* modality. The main differences in protocol between the two experiments is that during the presentation phase, participants were static and at 1.22 m from the screen in [24] whereas they were allowed to move in the exploration area and at 2.3 m or 3.3 m from the screen in the present experiment, see Fig. 3(a). Results for the *visual* modality and the results of [24] obtained in the real and virtual worlds are plotted in Fig. 8(a). From this figure, it can be seen that for *Position 1*, the results of the present experiment closely follow the real world results of [24] for $d_r < 3$ m and tend toward those for the virtual world when $d_r > 3$ m. For *Position 2*, the results of the present experiment closely follow real world results of [24] up to $d_r = 4.5$ m, before decreasing slightly. It can thus be concluded that moving during the *presentation* phase may have provided participants with a better visual distance perception for close distances. One can also notice that in the virtual world results taken from [24] the anchoring distance D_a (estimated here as the distance for which $d_r = d_p$) is around 1.4 m, and is close to 1.22 m, the distance between the participants and the screen. This also constitutes additional experimental evidence arguing in favor of the anchor hypothesis proposed in Sec. 4.2.

5.2 Perceived distance in *audio* VR-systems based on holophonic sound rendering

As discussed in Sec. 4.2, a compressive curve in the form of $d_p = k(d_r)^a$ has been shown to be a good model for the psychophysical function that relates estimates of perceived distance to physical source distance for the *audio* modality [42]. A review among 84 experiments is presented in [42] with a mean value of $a = 0.54$ obtained for the compression coefficient when fitting a compressive model to all the available data. It has also been observed that experimental protocols [26] (verbal report, perceptually directed action ...) and listening conditions [38] (static or moving) have very little influence on the obtained values of a.

Results from Sec. 3.2 for the *audio* condition are compared to this compressive model in Fig. 8(b). By fitting such a model to the perceived audio distances collected for *Position 1*, values of $a = 0.41 \pm 0.03$ and $k = 1.62 \pm 0.09$ are found, with $R^2 = 98\%$ of the variance observed in the experimental data explained

by the compressive model. The fit for *Position 2* gives values of $a = 0.29 \pm 0.07$ and $k = 2.13 \pm 0.31$, with $R^2 = 84\%$. The model $d_p = k(d_r)^a$ thus fits very well to the experimental data for both starting positions. The perception of auditory distance seems to be slightly more compressed in the virtual world than predicted in the real world using the average compressive model. However, a more rigorous experimental protocol, which compares directly real world and virtual world distance perception using the same distances and reporting method (as done in [23] for example), is needed to assess this point. It can nevertheless be concluded that WFS is able to synthesise sound-fields which are perceptually meaningful in terms of distance for *moving participants* and *static virtual* sources placed in the *action space*, apparently exhibiting slightly more compression than in the real world.

5.3 Utility of dynamic distance cues

It is important to notice that, as shown in Sec. 3.2, all of the participants spontaneously walked during the exploration phase and that the exploration durations t_{XP} were different among the different modalities, with $t_{XP}(A) > t_{XP}(AV) > t_{XP}(V)$. Participants thus attempted to gain information from the AV dynamic cues and seemed to proceed differently depending on the available modality. Moreover, they walked slightly more in the direction parallel to the virtual object than in the direction perpendicular to the virtual object during the presentation phase. This highlights the importance of dynamic cues in virtual audio-visual environments and provides some information concerning how perceptual cues may be weighted.

5.4 Feeling of presence and visual distance underestimation

The major problem related to the observed *presence* feeling is that it is participant-dependent. For some participants, *presence* was higher in the AV condition than in the V condition. For others, the opposite was true. This is potentially a consequence of the chosen audio stimulus (low pass filtered white noise, see Sec. 2.3) which was reported as unpleasant by some participants, and thus may have decreased their feeling of presence. This may also explain why no significant differences were found for the *presence*-score between the V condition and the AV conditions (see Fig. 6).

Furthermore, it has been suggested that because AV VR-systems provide a higher degree of *presence* than visual-only VR-systems, they potentially lead to less visual distance underestimation [22]. The correlation between the feeling of *presence* and visual distance underestimation is studied here. For each participant, the *presence* variation Δ_P induced by the addition of the spatialized audio stimuli is calculated as the difference between the *presence*-score of that participant in the AV condition and the *presence*-score of that participant in the V condition (see Sec. 3.4). Similarly, the *linear visual underestimation factor variation* Δ_α induced by the addition of the spatialized audio stimuli is calculated as the difference between the *linear underestimation factor* of that participant in the AV condition and the *linear underestimation factor* of that participant in the V condition. The *linear underestimation factor* is computed as the linear slope of the psychophysical curve relating d_p and d_r. As a result Δ_α and Δ_P are not found to be correlated (correlation coefficient of $\Gamma = -0.12$ and $p < 0.45$). Thus, this does not allow one to conclude that a higher degree of *presence* leads to less visual distance underestimation and illustrates the limited efficiency of post-session questionnaires as a tool to measure fine variations of presence.

6 Conclusion

In this paper a study of *audio*, *visual*, and *audio-visual* egocentric distance perception by moving participants in *virtual* environments is presented. Audio-visual rendering was provided by tracked passive visual stereoscopy and acoustic wave field synthesis (WFS). For each rendering condition, the estimation of perceived distances was based on a perceptually directed action using the method of indirect blind-walking. Distances perceived in the virtual environment were accurately estimated or overestimated for rendered distances closer than the audio-visual rendering system and underestimated for distances farther. Interestingly, participants perceived each virtual object at a modality-independent distance when using the audio modality, the visual modality, or the combination of both. Regarding the *audio* modality, WFS was able to synthesize perceptually meaningful sound-fields in terms of distance. Dynamic audio-visual cues are used by participants when estimating the distance of virtual objects. Moving may

have provided participants with a better visual distance perception of close distances than if they were static. No correlation between the feeling of *presence* and visual distance underestimation has been found. Finally, to explain the observed perceptual distance compression, it is proposed that, due to conflicting distance cues, the audio-visual rendering system physically *anchors* the virtual world to the real world, by attracting the virtual objects to it.

Acknowledgements

The authors wish to thank all the volunteers who took part in the experiment and *sonic emotion* for providing the *Wave 1* WFS rendering engine. Thomas Chartier and Philippe Cuvillier, now former students from the École Polytechnique (France), are also thanked for their help in the design and preliminary tests and Marc Fuzellier for his useful help during the second phase of the experiment. Special thanks are given to Matthieu Courgeon for time spent on the visual rendering. Finally, the authors would like to thank Antonio Trujillo-Ortiz for providing a reliable Matlab version of the repeated-measures two-way analysis of variance test.

References

[1] I. V. Alexandrova, P. T. Teneva, S. de la Rosa, U. Kloos, H. H. Bülthoff, and B. J. Mohler. Egocentric distance judgments in a large screen display immersive virtual environment. In *Proceedings of the 7th Symposium on Applied Perception in Graphics and Visualization*, APGV '10, pages 57–60, New York, NY, USA, 2010. ACM.

[2] J. Andre and S. Rogers. Using verbal and blind-walking distance estimates to investigate the two visual systems hypothesis. *Perception & Psychophysics*, 68(3):353–361, April 2006.

[3] C. Armbruster, M. Wolter, T. Kuhlen, W. Spijkers, and B. Fimm. Depth perception in virtual reality: Distance estimations in peri- and extrapersonal space. *Cyberpsychology & Behavior*, 11(1):9–15, February 2008.

[4] D. H. Ashmead, D. L. Davis, and A. Northington. Contribution of listeners approaching motion to auditory distance perception. *Journal of Experimental Psychology - Human Perception and Performance*, 21(2):239–256, April 1995.

[5] A. C. Beall, J. M. Loomis, and J. W. Philbeck. Absolute motion parallax weakly determines visual scale. *Investigative Ophthalmology & Visual Science*, 35(4):2111–2111, March 1994.

[6] A. J. Berkhout, D. de Vries, and P. Vogel. Acoustic control by wave field synthesis. *Journal of the Acoustical Society of America*, 93(5):2764–2778, 1993.

[7] J. Blauert. *Spatial Hearing, The Psychophysics of Human Sound Localization*. MIT Press, 1999.

[8] K. Bormann. Presence and the utility of audio spatialization. *Presence-Teleoperators and Virtual Environments*, 14(3):278–297, June 2005.

[9] A. W. Bronkhorst and T. Houtgast. Auditory distance perception in rooms. *Nature*, 397(6719):517–520, February 1999.

[10] F. P. Brooks. What's real about virtual reality? *IEEE Computer Graphics And Applications*, 19(6):16–27, 1999.

[11] E. Corteel. *Caractérisation et Extensions de la Wave Field Synthesis en conditions réelles*. PhD thesis, Université de Paris 6, 2004.

[12] E. Corteel, K. V. NGuyen, O. Warusfel, T. Caulkins, and R. Pellegrini. Objective and subjective comparison of electrodynamic and map loudspeakers for wave field synthesis. *30th International Conference of the Audio Engineering Society*, 2007.

[13] S. H. Creem-Regehr, P. Willemsen, A. A. Gooch, and W. B. Thompson. The influence of restricted viewing conditions on egocentric distance perception: Implications for real and virtual indoor environments. *Perception*, 34(2):191–204, 2005.

[14] N. Côté, V. Koehl, M. Paquier, and F. Devillers. Interaction between auditory and visual distance cues in virtual reality applications. In *Proceedings of the Forum Acusticum, Aalborg, Denmark*, 2011.

[15] J. E. Cutting. How the eye measures reality and virtual reality. *Behavior Research Methods Instruments & Computers*, 29(1):27–36, February 1997.

[16] R. R. A. Faria, M. K. Zuffo, and J. A. Zuffo. Improving spatial perception through sound field simulation in vr. *Proceedings of the 2005 IEEE International Conference on Virtual Environments, Human-Computer Interfaces and Measurement Systems*, pages 103–108, 2005.

[17] S. S. Fukusima, J. M. Loomis, and J. A. DaSilva. Visual perception of egocentric distance as assessed by triangulation. *Journal of Experimental Psychology - Human Perception and Performance*, 23(1):86–100, February 1997.

[18] Michael A. Gerzon. Ambisonics in multichannel broadcasting and video. *Journal of the Audio Engineering Society*, 33(11):859–871, 1985.

[19] T. Y. Grechkin, T. D. Nguyen, J. M. Plumert, J. F. Cremer, and J. K. Kearney. How does presentation method and measurement protocol affect distance estimation in real and virtual environments? *ACM Transactions on Applied Perception*, 7(4):26, July 2010.

[20] D. M. Hoffman, A. R. Girshick, K. Akeley, and M. S. Banks. Vergence-accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of Vision*, 8(3):33, 2008.

[21] P. A. Howarth. Potential hazards of viewing 3-D stereoscopic television, cinema and computer games: a review. *Ophthalmic and Physiological Optics*, 31(2):111–122, March 2011.

[22] V. Interrante, B. Ries, J. Lindquist, M. Kaeding, and L. Anderson. Elucidating factors that can facilitate veridical spatial perception in immersive virtual environments. *Presence-Teleoperators and Virtual Environments*, 17(2):176–198, April 2008.

[23] Gavin Kearney, Marcin Gorzel, Henry Rice, and Frank Boland. Distance perception in interactive virtual acoustic environments using first and higher order ambisonic sound fields. *Acta Acustica united with Acustica*, 98(1):61–71, 2012.

[24] E. Klein, J. E. Swan, G. S. Schmidt, M. A. Livingston, and O. G. Staadt. Measurement protocols for medium-field distance perception in large-screen immersive displays. *IEEE Virtual Reality 2009, Proceedings*, pages 107–113, 2009.

[25] Setsu Komiyama, Akira Morita, Kohichi Kurozumi, and Katsumi Nakabayashi. Distance control system for a sound image. In *Audio Engineering Society Conference: 9th International Conference: Television Sound Today and Tomorrow*, 2 1991.

[26] J. M. Loomis, R. L. Klatzky, J. W. Philbeck, and R. G. Golledge. Assessing auditory distance perception using perceptually directed action. *Perception & Psychophysics*, 60(6):966–980, 1998.

[27] J. M. Loomis and J. M. Knapp. *Virtual and Adaptive Environments: Applications, Implications, and Human Performance Issues*. Lawrence Erlbaum Associates, 2003.

[28] A. Naceri, R. Chellali, F. Dionnet, and S. Toma. Depth perception within virtual environments: a comparative study between wide screen stereoscopic displays and head mounted devices. *2009 Computation World: Future Computing, Service Computation, Cognitive, Adaptive, Content, Patterns*, pages 460–466, 2009.

[29] M. Nawrot and K. Stroyan. The motion/pursuit law for visual depth perception from motion parallax. *Vision Research*, 49(15):1969–1978, July 2009.

[30] Jodie M. Plumert, Joseph K. Kearney, James F. Cremer, and Kara Recker. Distance perception in real and virtual environments. *ACM Transactions on Applied Perception*, 2:216–233, July 2005.

[31] C. Porschmann and C. Storig. Investigations into the velocity and distance perception of moving sound sources. *Acta Acustica United With Acustica*, 95(4):696–706, July 2009.

[32] M. Rébillat, E. Corteel, and B. F.G. Katz. SMART-I^2 "Spatial Multi-user Audio-visual Real-Time Interactive Interface". *125th Convention of the Audio Engineering Society*, 2008.

[33] M. Rébillat, B. F.G. Katz, and E. Corteel. SMART-I^2: "Spatial Multi-user Audio-visual Real-Time Interactive Interface", a broadcast application context. *Proceedings of the IEEE 3D-TV conference*, 2009.

[34] M. K. Russell and A. L. Schneider. Sound source perception in a two-dimensional setting: Comparison of action and nonaction-based response tasks. *Ecological Psychology*, 18(3):223–237, 2006.

[35] J. Ryu, N. Hashimoto, and M. Sato. Influence of resolution degradation on distance estimation in virtual space displaying static and dynamic image. *2005 International Conference on Cyberworlds, Proceedings*, pages 43–50, 2005.

[36] J. Sanson, E. Corteel, and O. Warusfel. Objective and subjective analysis of localisation accuracy in wave field synthesis. *124th Convention of the Audio Engineering Society*, 2008.

[37] M. Slater and S. Wilbur. A framework for immersive virtual environments (five): Speculations on the role of presence in virtual environments. *Presence-Teleoperators and Virtual Environments*, 6(6):603–616, December 1997.

[38] J. M. Speigle and J. M. Loomis. Auditory distance perception by translating observers. *Proceedings of IEEE Symposium on Research Frontiers in Virtual Reality, San Jose, CA, October 25-26.*, 1993.

[39] J. R. Springer, C. Sladeczek, M. Scheffler, J. Hochstrate, F. Melchior, and B. Frohlich. Combining wave field synthesis and multi-viewer stereo displays. *IEEE Virtual Reality 2006, Proceedings*, pages 237–+, 2006.

[40] S. J. Watt, K. Akeley, M. O. Ernst, and M. S. Banks. Focus cues affect perceived depth. *Journal of Vision*, 5(10):834–862, 2005.

[41] W. M. Wiest and B. Bell. Stevens exponent for psychophysical scaling of perceived, remembered, and inferred distance. *Psychological Bulletin*, 98(3):457–470, 1985.

[42] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst. Auditory distance perception in humans: A summary of past and present research. *Acta Acustica United With Acustica*, 91(3):409–420, May 2005.